Multi-Metric Evaluation of Large Language Models for Legislative Consolidation of Italian Senate Bills

Simone Vagnoni
CIRSFID ALMA-AI,
University of Bologna
Bologna, Italy
simone.vagnoni3@unibo.it

Michele Corazza
CIRSFID ALMA-AI,
University of Bologna
Bologna, Italy
michele.corazza2@unibo.it

Monica Palmirani
CIRSFID ALMA-AI,
University of Bologna
Bologna, Italy
monica.palmirani@unibo.it

ABSTRACT

Legislative text consolidation is a labour-intensive task and requires legal expertise, semantic understanding of the context, and structural precision. The main goal is to apply amendments submitted by members of the Parliament concerning a proposed draft legislation in order to obtain an updated version of the bill after the assembly's vote. We evaluate six frontier LLMs (Claude Sonnet 4.5/3.5, GPT-5.1 Codex, Gemini 3 Flash, Kimi K2.5, DeepSeek v3.2) on 67 Italian Senate bills (23,416 approved amendments) with official ground truth from the Senate Akoma Ntoso bulk repository [1] in this task. We report five complementary metrics spanning content fidelity (L0, BLEU-2, ROUGE-2), semantic preservation (L1), and structural integrity (L2).

Results on complete outputs within the common attempted set (67 bills) show that Claude Sonnet 4.5 achieves L0=0.769, L1=0.936, L2=0.990, while GPT-5.1 Codex attains L1=0.870 with L0=0.424. DeepSeek v3.2 achieves L2=0.873 yet L0=0.210. The key finding is a decoupling between structural validity and content fidelity: high L2 does not guarantee high L0. This implies evaluation of constrained LLM generation must assess multiple dimensions. We release methodology, metrics, prompt templates, and an error taxonomy for replication.

CCS CONCEPTS

• **Applied computing** → **Law**; • **Computing methodologies** → **Natural language processing**; • **Information systems** → **Information retrieval**.

KEYWORDS

legislative consolidation, large language models, Akoma Ntoso, evaluation metrics, legal AI

1 INTRODUCTION

Legislative consolidation is the process of integrating approved parliamentary amendments into the proposed draft legislative document[15][16]. In the Assembly of Italian Senate, this task transforms a bill and its approved amendments into an official report (*A package* consolidated bill coming from the Committee and ready for Assembly or *DDL Messaggio* consolidated bill ready for the Chamber of Deputies) the authoritative text for subsequent parliamentary proceedings. This task requires a specific legal expert competence, especially for some amendments that are complex and have consequential changes in the rest of the document (cascade amendment) or modifications to existing amendments (subamendment). According to the World e-Parliament Report 2024 [13], more than two-thirds of the 115 parliamentary chambers surveyed across 86 countries have already adopted multi-year digital strategies that include artificial intelligence, especially addressed to the amendment management.

The Italian Senate has been using an amendment management suite called GEM to group similar texts and to order them semi-automatically. GEM does not perform the consolidation task. The aim of this paper is to evaluate several large language models (LLMs), using targeted prompts, to carry out this consolidation [14]. The input consists of the official bills and amendments provided by the Italian Senate in Akoma Ntoso format, while the reference output is the consolidated text produced by high-level expert human drafters. Our goal requires interpreting natural language amendment instructions, locating precise targets within hierarchical document structures, and regenerating XML-compliant output conforming to the Akoma Ntoso standard [1].

Prior approaches to automation have relied on pattern matching and rule extraction [2]. The emergence of Transformer-based large language models with extended context windows (128K–1M tokens) suggests potential for broader automation [10]. However, legislative consolidation presents distinct challenges: unlike comprehension tasks evaluated in prior legal AI work [3–5], consolidation requires generating structured XML that must satisfy both semantic correctness and formal schema constraints.

This paper addresses three research questions:

RQ1 How do frontier LLMs perform on legislative consolidation across multiple evaluation dimensions?
RQ2 What failure modes characterize LLM-generated legislative XML?

[1] https://github.com/SenatoDellaRepubblica/AkomaNtosoBulkData

RQ3 Is structural validity sufficient to ensure consolidation quality?

1.1 Contributions

This work makes three contributions:

(1) **Multi-metric evaluation framework**: We define five complementary metrics—Levenshtein similarity (L0), BLEU-2, ROUGE-2, SBERT-based semantic similarity (L1), and a weighted structural integrity score (L2)—capturing quality dimensions no single metric addresses.
(2) **Full-document LLM benchmark in Akoma Ntoso**: We establish the first full-document benchmark for LLM-based legislative consolidation in Akoma Ntoso, covering 67 Italian Senate bills (23,416 amendments) with official consolidated texts as ground truth. This extends prior article-level work [12] to full-document scope, revealing scale-dependent failure patterns invisible at article granularity. Coverage statistics are computed over the common attempted set (including failures), while metric aggregates use complete outputs only.
(3) **Structural validity is not sufficient**: We demonstrate that a model can achieve acceptable structural validity (high L2) while producing low-fidelity content (low L0), establishing that multi-metric evaluation is necessary.

We also provide an error taxonomy of failure modes across models and release the complete prompt template, per-case metrics, and evaluation code for replication.

2 BACKGROUND

2.1 Automatise the Legislative Bill Consolidation

Legislative consolidation integrates approved amendments into a base bill to produce an authoritative consolidated text. In the Italian Senate, this transforms a *DDL Commissione* (committee bill) into a report A, ready for the Assembly step, and the Assembly voted output into another consolidated text called *DDL Messaggio* (consolidated bill). Computationally, the task requires: (1) parsing amendment directives expressed in natural language, (2) locating precise targets within a hierarchical document, (3) applying edits (substitution, insertion, deletion), and (4) regenerating a valid XML document. Linguistic variation makes steps (1)–(2) difficult: the same operation may appear in many forms (e.g., "*Article 3 is replaced by the following:*" vs. "*Replace Article 3 with the following:*"). The output must preserve legal meaning and structure, not just textual content. There are also complex situations like the consequential amendments where a modification generates in another part of the bill other modifications (e.g., "*In paragraph 755, replace the words "100 million euros" with "150 million euros." Consequently, delete paragraph 764.* "). Another hard task is due to modifications of the amendments called sub-amendment (e.g,"*To amendment 18.55 (Text 2), after "Consequently," and the related changes to paragraph 576, add the following modifications*"). These kinds of amendments are not in the scope of this paper.

2.2 Structural Constraints of the Akoma Ntoso Specification

The Italian Senate publishes legislative documents in Akoma Ntoso format, an OASIS standard for parliamentary XML [1]. A valid consolidated document must satisfy multiple structural constraints:

Hierarchical structure: Documents follow a strict nesting hierarchy. The root `<akomaNtoso>` element contains a `<bill>` or `<act>` document type. Within this, `<meta>` holds identification and lifecycle metadata, while `<body>` contains the legislative content organized as `<article>` elements, each containing `<paragraph>` elements, potentially with `<point>` and `<letter>` subdivisions. Each structural element carries a unique `eId` attribute (e.g., `eId="art_1-par_2"`) enabling precise cross-references.

FRBR metadata: The Functional Requirements for Bibliographic Records model requires identification at three levels: FRBRWork (the abstract legislative concept), FRBRExpression (its Italian-language expression and versions over time), and FRBRManifestation (this specific XML file). Each level contains URI identifiers, dates, and authorship references.

Cross-reference integrity: Internal references (`<refhref="#art_3-par_1">`) must resolve to existing `eId` targets. Broken references—pointing to elements that don't exist—compromise legal interpretation when readers encounter phrases like "as defined in paragraph 1" with no valid target.

Italian numbering conventions: Article insertions follow the pattern 1, 2, 2-bis, 2-ter, 2-quater, 3—preserving original numbering for citation stability while accommodating insertions. Paragraph numbering follows similar conventions within each article.

Metadata: Amendments and Consolidated text require metadata of consolidation that are recorded in the `<analysis>` elements respectively in `<activeModifications>` for the amendment and in the `<passiveModifications>` for the updated text. Sub-elements of both store many relevant information like the type of change, destination, source, and

content. Fortunately, the amendments are timeless operations, so we don't have in this scenario the complication of the temporal aspects (e.g., postponement of the modification).

Version of AKN schema: The Italian Senate uses the CSD03 namespace with custom extensions that differ from the official OASIS AKN schema. Documents that pass Senate validation may fail strict XSD checks because CSD03 permits `id` attributes where the new approved OASIS XSD forbids them. This requires relaxed validation approaches that verify essential structure without rejecting valid parliamentary documents.

2.3 Research Gap and Prior Approaches

Prior legal AI evaluation in the community of AI and Law has focused on comprehension tasks. Some examples are the bar examination performance [4], contract clause extraction [3], and legal question answering [5]. These assess whether models *understand* legal text through classification or extraction.

Legislative consolidation differs fundamentally because the models must *generate* structured output satisfying both semantic and syntactic constraints simultaneously. This is closer to code generation than text comprehension—the output must parse correctly (syntactic validity) while implementing correct behavior (semantic correctness).

Rule-based consolidation systems achieve limited coverage due to linguistic variation in amendment phrasing [2]. Pattern matching handles common phrasings but fails on paraphrases, nested operations, or implicit references. Recent work by Etcheverry et al. [12] addresses consolidation at the *single-article* level for French tax law using triplets (original article, modification section, modified article) and Word Error Rate. Our setting differs in scope and constraints because we generate full Akoma Ntoso XML documents and evaluate semantic and structural validity in addition to textual fidelity. Methodologically, these tasks are not directly comparable; our contribution is a multi-metric framework for constrained generation at full-document scale.

3 EVALUATION FRAMEWORK

We define five metrics capturing complementary quality dimensions. The rationale for multiple metrics is that each addresses failure modes invisible to others. We present character-level fidelity (L0), bigram precision (BLEU-2), bigram overlap (ROUGE-2), semantic preservation (L1), and structural integrity (L2).

Table 1: Metric Selection Rationale

Metric	Dimension	What it measures	Limitation
L0 (Lev.)	Content fidelity	Character-level match	Ignores paraphrases
BLEU-2	Content fidelity	Bigram precision	Limited recall
ROUGE-2	Content fidelity	Bigram overlap/recall	Not structural
L1 (Sem.)	Semantic preservation	Meaning retention	Blind to structure
L2 (Struct.)	Structural validity	Multi-component validity	Ignores content

3.1 Layer 0: Character-Level Fidelity (Levenshtein)

We extract text content from XML (stripping tags) and compute normalized Levenshtein similarity [17]:

$$L_0(D_c, D_r) = 1 - \frac{\text{Lev}(\text{text}(D_c), \text{text}(D_r))}{\max(|\text{text}(D_c)|, |\text{text}(D_r)|)} \quad (1)$$

where D_c is the consolidated output, D_r is the reference document, $\text{text}(\cdot)$ extracts concatenated text content, and $\text{Lev}(\cdot,\cdot)$ computes the Levenshtein edit distance (minimum number of single-character insertions, deletions, or substitutions required to transform one string into another).

3.2 BLEU-2 and ROUGE-2: Bigram Similarity

We compute bigram metrics over extracted plain text. BLEU-2 measures modified bigram precision [8], while ROUGE-2 measures bigram recall/overlap [9]. Both use the same normalized text as L0.

Interpretation: ROUGE > BLEU indicates over-generation (extra content beyond reference), while BLEU > ROUGE indicates omissions. These metrics help distinguish "correct but verbose" outputs from "terse but incomplete" ones.

3.3 Layer 1: Semantic Preservation

Character-level and n-gram metrics share a fundamental limitation: they penalize semantically equivalent paraphrases. In legislative text, "Il Ministro dell'Economia e delle Finanze" and "Il Ministro dell'economia e delle finanze" differ only in capitalization but are legally identical. More significantly, "entro il termine di trenta giorni" (within the term of thirty days) and "entro trenta giorni" (within thirty days) are semantically equivalent despite different surface forms.

We define L1 to measure meaning preservation independent of lexical variation, using Sentence-BERT [6, 7].

Documents are segmented into sentences using the following procedure:

(1) Extract text content from each <p> element in the XML body
(2) Apply Italian sentence boundary detection (splitting on "." followed by whitespace and uppercase letter, with exceptions for common abbreviations: "art.", "n.", "co.", "lett.")
(3) Filter sentences shorter than 10 characters (typically numbering or formatting artifacts)

Let $S_r = \{s_r^1, s_r^2, \ldots, s_r^m\}$ denote sentences from the reference document and $S_c = \{s_c^1, s_c^2, \ldots, s_c^n\}$ sentences from the consolidated output.

We use `paraphrase-multilingual-MiniLM-L12-v2`, a multilingual model that allows us to encode sentences into dense vectors. These dense vectors can be compared using cosine similarity to produce a proxy for semantic similarity. We selected this model for its strong multilingual coverage (including Italian), paraphrase-oriented training, and low inference cost suitable for batch evaluation at scale.

We compute L1 as the average maximum similarity between each reference sentence and its best match in the generated output:

$$L_1(D_c, D_r) = \frac{1}{|S_r|} \sum_{s_r \in S_r} \max_{s_c \in S_c} \text{sim}(s_r, s_c) \tag{2}$$

where similarity uses cosine distance:

$$\text{sim}(s_r, s_c) = \frac{\text{emb}(s_r) \cdot \text{emb}(s_c)}{\|\text{emb}(s_r)\| \cdot \|\text{emb}(s_c)\|} \tag{3}$$

We iterate over reference sentences and find their best matches in the generated output, not vice versa. This design choice reflects our evaluation goal: measuring how completely the model *preserves* reference content. A model that includes all reference content plus additional material achieves L1=1.0—appropriate because the reference content is preserved. The converse (iterating over generated sentences) would penalize models for including any content not in the reference, conflating completeness with minimality.

While L1 captures content presence, it cannot consider whether the content appears in the correct structure. For this reason we develop L2, which considers the structural integrity of the documents.

3.4 Layer 2: Structural Integrity

Valid Akoma Ntoso requires more than well-formed XML. We define L2 as a weighted composite over four orthogonal structural dimensions to reflect functional validity under the Italian CSD03 practice.

$$L_2 = 0.4 \cdot V_{\text{schema}} + 0.3 \cdot V_{\text{refs}} + 0.2 \cdot V_{\text{num}} + 0.1 \cdot V_{\text{meta}} \tag{4}$$

V_{schema} (40%): Essential Element Presence. We verify presence of required Akoma Ntoso structural elements:

- Root element: <akomaNtoso> or namespaced <an:akomaNtoso>
- Document type: <bill> or <act> as direct child
- Metadata section: <meta> element present
- Content section: <body> element present
- Legislative content: At least one <article> element within body

V_{schema} is a graded score: each of the five checks contributes 0/1, and V_{schema} is their average. This preserves diagnostic detail while keeping a single component score.

V_{refs} (30%): Cross-Reference Integrity. Legislative documents contain internal references ("as defined in Article 3, paragraph 1"). In Akoma Ntoso, these are encoded as:

```
<ref href="#art_3-par_1">paragraph 1</ref>
```

We extract all `href` attributes beginning with "#" (internal references) and verify each target `eId` exists in the document:

$$V_{\text{refs}} = \frac{|\{r \in R : \text{target}(r) \in E\}|}{|R|} \tag{5}$$

where R is the set of internal references and E is the set of all `eId` values in the document. If $|R| = 0$ (no internal references), $V_{\text{refs}} = 1.0$.

V_{num} (20%): Numbering Sequence Correctness. Italian legislative numbering follows specific conventions that must be preserved for citation stability:

- Articles: 1, 2, 2-bis, 2-ter, 2-quater, 3, 4, ...
- Paragraphs within articles: 1, 2, 2-bis, 3, ...
- The "-bis", "-ter", "-quater" suffixes indicate inserted elements

We verify:

(1) No duplicate article or paragraph numbers within scope
(2) Sequences are monotonically increasing (accounting for Latin suffixes)
(3) Format matches pattern: digit(s) optionally followed by "-bis", "-ter", etc.

V_{num} is computed as:

$$V_{\text{num}} = 1 - \frac{\text{violations}}{\text{total elements}} \tag{6}$$

where violations include duplicates, gaps, and format errors.

V_{meta} (10%): FRBR Metadata Completeness. The Functional Requirements for Bibliographic Records (FRBR) model requires identification at three abstraction levels. We verify presence of:

- <FRBRWork> with <FRBRthis>, <FRBRuri>, <FRBRdate>
- <FRBRExpression> with <FRBRthis>, <FRBRuri>, <FRBRdate>
- <FRBRManifestation> with <FRBRthis>, <FRBRuri>, <FRBRdate>

V_{meta} is the fraction of required elements present (9 total).

In our validation tests, strict OASIS WD17 XSD checks reject Italian Senate CSD03 documents (e.g., `id` attributes in metadata). We therefore use functional checks with diagnostic granularity and different weights to account for the importance of various structural aspects.

4 EXPERIMENTAL SETUP

4.1 Dataset Construction

We collected data from the Italian Senate Akoma Ntoso bulk repository.[2] Selection criteria required complete data availability: base bill (*DDL Commissione*), approved amendment list with full text, and official consolidated output (*DDL Messaggio*) published after committee approval. We exclude budget/tabular (BGT) cases and documents whose consolidated output is table-only (no articles), to avoid non-textual formats.

From the 17th Legislature (2013–2018; 66/67 cases) we identified 67 bills that satisfy these criteria and were attempted by all models in the common evaluation set; one case belongs to the 18th Legislature. The dataset comprises 23,416 approved amendments. We exclude cases lacking any of the three components (base bill, amendments, or official DDL Messaggio), as well as table-only consolidated outputs. This filtering reflects practical availability and the need for text-based evaluation.

Table 2: Dataset Summary (67 Bills)

	Amendments	Bills
Total	23,416	67
Min/Max per bill	1 / 7,405	—
Median per bill	83	—
Mean per bill	349.5	—

Complexity classification: We stratify cases by amendment count and document length for analysis. Full case lists and per-case metadata are available in the released manifest (`results/experiments/selected_cases.json`).

4.2 Models and Configuration

We evaluate six frontier LLMs representing major providers and architectural approaches:

[2] https://github.com/SenatoDellaRepubblica/AkomaNtosoBulkData/

Table 3: Evaluated Models and Specifications

Model	Provider	Context	Access
Claude Sonnet 4.5	Anthropic	1M	OpenRouter
Claude 3.5 Sonnet	Anthropic	200K	OpenRouter
GPT-5.1 Codex Max	OpenAI	400K	OpenRouter
Gemini 3 Flash	Google	1M	OpenRouter
Kimi K2.5	Moonshot AI	256K	OpenRouter
DeepSeek v3.2	DeepSeek	163K	OpenRouter

All models were accessed through OpenRouter's unified API, ensuring consistent request formatting, authentication, and response handling across providers. This eliminates potential confounds from provider-specific API behaviors.

We used identical parameters for all models: `temperature=0.0` (deterministic/greedy sampling), `max_tokens=64000`. Context windows reported are as available at time of evaluation (January 2026); a uniform max_tokens cap was applied across models. Temperature 0.0 targets the model's single best response rather than sampling variation.

4.3 Prompt Design

Prompt design was iteratively refined based on observed failures (missing <body>, invalid metadata attributes, and namespace inconsistencies). The final 12-rule template below enforces XML-only output, structural fidelity, and metadata constraints while preserving the base document structure.

Listing 1: Complete Consolidation Prompt Template

```
You are an expert legislative drafter. Produce
     the CONSOLIDATED Akoma Ntoso document by
    applying all approved amendments to the
    base DDL.

Mandatory rules:
1) Output ONLY a well-formed Akoma Ntoso XML
    document:  root must be <akomaNtoso> or <
    an:akomaNtoso> and must close. Do not wrap
     in Markdown or add comments/text outside
    XML.
2) Preserve the original structure (articles,
   paragraphs, letters) and modify only the
   targeted parts.
3) Apply ALL amendments in the order provided.
     None may be skipped.
4) Structural changes:
   - New paragraph: add <paragraph> with
       correct eId/numbering
   (e.g., 1-bis, 2-ter).
   - Full article replacement/insertion:
       update or insert <article> with
       coherent numbering.
5) Text changes:
   - Replace EXACT text/structures indicated
```

```
    (use <quotedStructure>/<quotedText>
        guidance).
    - Preserve identifiers/attributes (eId, wId
        , GUID) on elements inside the <body>.
6) Do NOT change articles/paragraphs not
    referenced by amendments.
No notes or free-text.
7) Ensure the output contains <body> with
    consolidated articles/paragraphs (not only
     <meta>).
8) If you cannot produce valid Akoma Ntoso,
    return an empty string.
9) Metadata update: In the <identification>
    block, set <FRBRsubtype> to "DDLMESS" and
    update <FRBRdate> date attribute to today'
    s date.
10) Preamble: Preserve the <preamble> block
    from the base DDL, if it exists.
11) Cover Page: Generate a minimal <coverPage>
     containing only a <p> with the document
    title.
12) Attribute Rules (Crucial for Validation):
    - The document element (e.g., <an:bill>)
        MUST have a name attribute (e.g., name
        ="bill").
    - Metadata elements inside <meta> (e.g., <
        componentData>, <eventRef>, <original
        >) MUST NOT use the id attribute. Use
        eId only for referentiable elements
        inside <body>.

Base DDL (Akoma Ntoso):
```xml
{base_xml}
```

Approved amendments to apply
({n_amendments} total): {amendment_list}

Final instruction: Return ONLY the final
    consolidated Akoma Ntoso XML, fully closed
    , with updated <body>.
No additional text.
```

Rules 7 and 12 directly address the most persistent failure modes: body omission and forbidden metadata attributes.

4.4 Execution Pipeline

We executed the pipeline on 67 bills in the common attempted set (intersection across models). Attempted runs are defined by the presence of a model run folder with at least one artifact (prompt, raw output, error log, or XML). We report results on successful runs (XML with all five metrics) per model (Table 5).

(1) **Input preparation**: Extract base bill XML from Senate archive; parse amendment list from accompanying JSON metadata; construct prompt by template substitution.
(2) **API invocation**: Submit prompt to model via OpenRouter; record response and timing; implement exponential backoff for rate limits.
(3) **Response extraction**: Strip Markdown code fences if present; detect and handle encoding issues (BOM, character escaping); validate XML well-formedness.
(4) **Text normalization**: For L0/BLEU-2/ROUGE-2 computation, extract text content from both generated and reference XML; normalize whitespace (collapse runs, trim); apply consistent Unicode normalization (NFC).
(5) **Metric computation**: Compute all five metrics against ground truth; record component scores for L2 diagnostics.

Several large bills (high amendment counts and large base XML) triggered timeouts or context-limit failures across models, leading to truncated or empty outputs. We keep these attempts in coverage tables and mark them as failures in the figures.

Table 4: Execution Coverage and Completeness (67 Bills, common attempted set)

Model	Att.	Comp.	Incomp.	Fail %
Claude Sonnet 4.5	67	45	22	32.8
Claude 3.5 Sonnet	67	26	41	61.2
GPT-5.1 Codex	67	37	30	44.8
Gemini 3 Flash	67	50	17	25.4
Kimi K2.5	67	33	34	50.7
DeepSeek v3.2	67	15	52	77.6
Total	**402**	**206**	**196**	**48.8**

In Table 4, **Comp.** denotes XML outputs. Incomplete outputs include missing XML (OpenRouter HTTP 400/403/402 dominate) and partial metrics where text-based scores (L0/L1) could not be computed. Across all incomplete runs, HTTP error 400 accounts for 41.3%, error 403 for 20.8%, error 402 for 7.7%, partial-metric cases for 23.9%, with the remainder split among empty responses and other errors. These are retained in coverage and figures.

5 RESULTS

5.1 Overall Performance (RQ1)

Table 5 presents aggregate results across complete outputs only (XML with all five metrics). We report means over the available cases per model.

Because some XML outputs lack text-based metrics (L0/L1), the effective sample size n in Table 5 can be smaller than the XML counts in Table 4.

In the table 5, n denotes outputs with all five metrics available (L0, BLEU-2, ROUGE-2, L1, L2). For Claude Sonnet 4.5,

Table 5: Performance Across Five Evaluation Metrics (Mean Scores on Complete Outputs)

Model	L0	BLEU-2	ROUGE-2	L1	L2	n
Claude Sonnet 4.5	**0.769**	**0.763**	**0.800**	**0.936**	0.990	44
Claude 3.5 Sonnet	0.632	0.566	0.608	0.877	0.979	19
Gemini 3 Flash	0.630	0.609	0.666	0.869	0.983	35
GPT-5.1 Codex	0.424	0.431	0.488	0.870	0.934	24
Kimi K2.5	0.436	0.439	0.515	0.788	**0.997**	18
DeepSeek v3.2	0.210	0.190	0.230	0.870	0.873	11

1 XML output is excluded due to a parsing failure (invalid QName).

In terms of L0, Claude Sonnet 4.5 leads on character-level and bigram metrics. The gap between Claude 4.5 (L0 0.769) and DeepSeek v3.2 (L0 0.210) indicates qualitative differences in output correctness. When considering L1, Claude 4.5 again shows the highest semantic preservation (L1 0.936), indicating strong meaning retention alongside high fidelity. Finally, in terms of L2 Kimi K2.5 achieves the highest structural score (L2 0.997), but structural validity alone does not guarantee content correctness. Figure 1 summarizes success rates by difficulty.

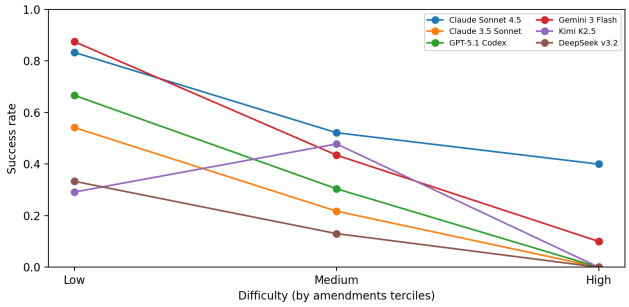

Figure 1: Success rate by difficulty (terciles of amendment count) for each model. Success rate is the fraction of attempted runs that produced a complete output (XML with all five metrics). Figure generated with Python: Pandas/Matplotlib/Seaborn.

5.2 ROUGE > BLEU Pattern

Across all models, ROUGE-2 exceeds BLEU-2 (e.g., Claude 4.5: 0.800 vs. 0.763; Gemini 3 Flash: 0.666 vs. 0.609), indicating systematic *over-generation*: models include most reference bigrams but add extra phrasing or formatting. In deployment, this error profile is preferable to omission, because superfluous text can be removed deterministically, whereas missing content requires re-analysis.

5.3 Analysis of Metric Patterns

For Claude Sonnet 3.5, we observe high values across L0/L1/L2, which indicate both semantic and structural correctness, with limited over-generation (ROUGE-2 slightly above BLEU-2).

GPT 5.1 Codex Max produces relatively high L1 but lower L0, which suggests meaning retention with lexical deviation, reducing character-level fidelity despite semantic alignment.

Gemini 3 Flash combines relatively high L2 with moderate L0, indicating consistent XML structure with some content loss or rephrasing.

Kimi K2.5 achieves the highest L2 (0.997) but only mid L0 (0.436) and lower L1 (0.788), underscoring that structural validity can be decoupled from content fidelity.

The final model, DeepSeek 3.2, produces moderate L2 paired with very low L0, reflecting task-level misalignment: structurally valid outputs can still contain incorrect content.

5.4 Model Robustness by Document Size

Across all models we observe that failures concentrate in larger bills. For DeepSeek v3.2, successful runs average 29.8 amendments, while failed runs average 441.7 (77.6% failure rate overall); for Kimi K2.5, 43.3 vs. 646.7; for Claude 4.5, 80.0 vs. 900.6. This suggests that practical context/latency limits disproportionately affect large documents even within their context windows.

5.5 Structural Validity Is Not Sufficient (RQ3)

Structural validity (L2) and content fidelity (L0) are weakly coupled in practice. A document can be well-formed and structurally complete yet implement the wrong semantic operation. This motivates multi-dimensional evaluation: validation should combine structural and content metrics rather than relying on structure alone.

6 FAILURE MODE ANALYSIS (RQ2)

We analyze 206 XML outputs across the six models using an automated structural analysis pipeline. Each output is classified into a single structural category by checking, in priority order: (1) XML well-formedness (parsing with `lxml`),

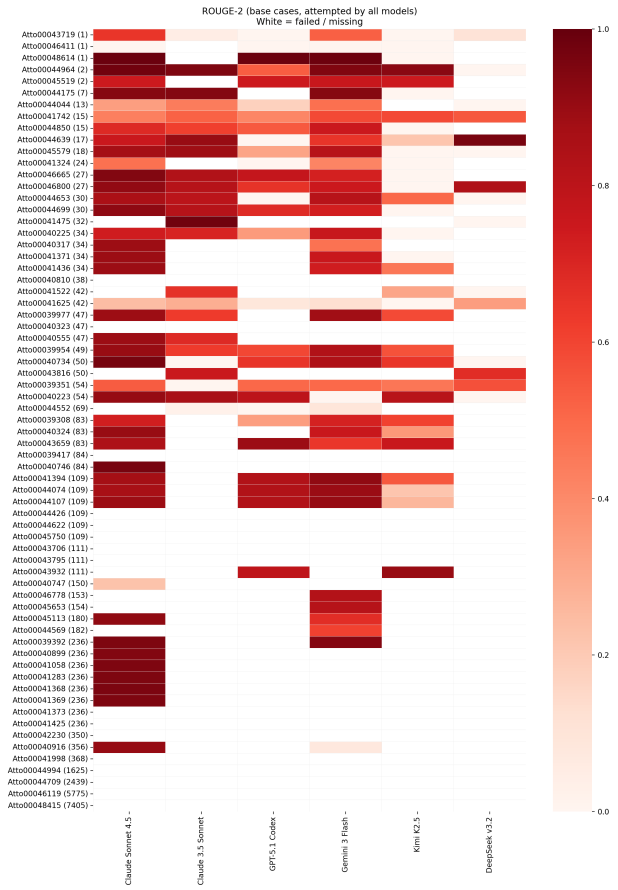

Figure 2: ROUGE-2 heatmap on the common attempted set (white = failed / missing). Figure generated with Python: Pandas/Matplotlib/Seaborn.

(2) presence of <body> element, (3) presence of at least one <article> element, and (4) whether all articles carry eId attributes. We also verify the presence of an <akomaNtoso> root; no parsed outputs failed this check, so it does not appear as a separate category. Table 6 summarizes the taxonomy with a per-model breakdown.

Table 6: Structural Error Taxonomy by Model (206 XML Outputs; C4.5/C3.5 = Claude Sonnet 4.5/3.5).

Error Type	C4.5	C3.5	GPT-5.1	Gemini	Kimi	DeepSeek
Perfect structure	42	20	6	45	20	6
Missing eId	1	6	23	4	0	2
Missing <body>	0	0	1	0	11	0
No <article>	1	0	7	1	1	7
Invalid XML	1	0	0	0	1	0

Example (missing eId on <article>). The snippet below shows an article with an id but no eId, which triggers the *missing* eId category.

```
<an:body>
  <an:article id="art1">
    <an:num>Art. 1.</an:num>
    <an:heading>(Candidabilita a cariche
        elettive...)</an:heading>
    <an:paragraph id="art1-par1">
      <an:num>1.</an:num>
      <an:content>
```

7 DISCUSSION

7.1 Implications for Evaluation Methodology

Our results establish that single-metric evaluation is insufficient for legal document generation. Each metric captures distinct failure modes while remaining blind to others:

- **L0** captures character-level differences but conflates substantive errors with formatting variations and misses word-order changes.
- **BLEU-2/ROUGE-2** add word-order sensitivity and precision/recall balance, but remain blind to XML structure.
- **L1** measures semantic preservation but is blind to structure; content can appear in the wrong location.
- **L2** verifies formal validity but says nothing about content correctness—structural validity alone is not sufficient.

The DeepSeek finding—acceptable L2 with catastrophic L0—demonstrates that production systems must evaluate multiple complementary dimensions. We recommend minimum thresholds on *all* metrics rather than any single gating criterion.

7.2 Deployment Architecture

Based on our findings, we recommend a four-stage pipeline:

(1) **LLM generation**: Claude Sonnet 4.5 produces initial consolidation.
(2) **Multi-metric validation**: Automated checks compute L0, BLEU-2, ROUGE-2, L1, and L2. Flag outputs failing any threshold (suggested: L0 < 0.5, L2 < 0.9, any L2 component < 0.6).
(3) **Deterministic recovery**: Auto-correct predictable errors—regenerate missing eId attributes from document structure, fix metadata patterns, restore formatting elements.
(4) **Expert review**: Human drafter verifies legal correctness for flagged cases and approves final output.

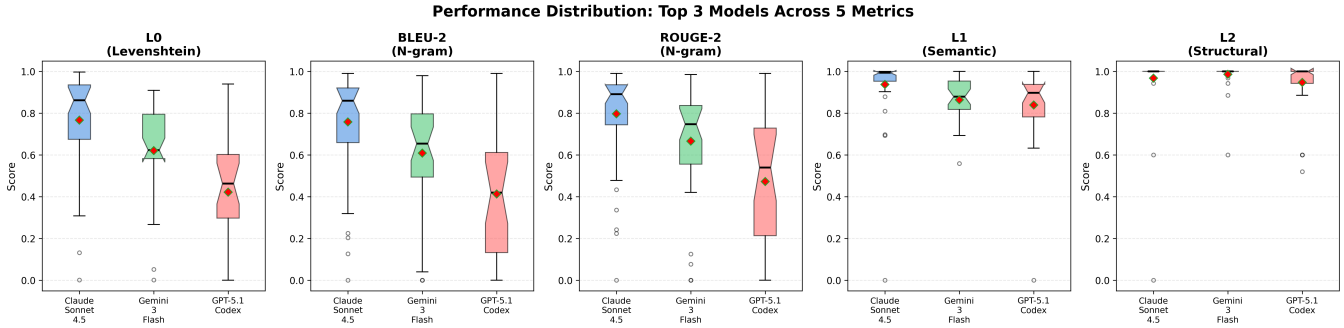

Figure 3: Top-3 model score distributions (L0/L1/L2) on the test set. Fig. generated with Python: Matplotlib.

This architecture reduces manual effort while maintaining human oversight for legal accountability.

7.3 Limitation in the metrics

In this study, we propose metrics to measure character-level match (L0), bigram-based fidelity (BLEU-2/ROUGE-2), semantic preservation (L1), and structural validity (L2).

Some crucial criteria are however still lacking, such as independent legal correctness beyond the official *DDL Messaggio*, generalization to other legislatures or languages, robustness to alternative prompts, or long-term model stability beyond the January 2026 snapshot. These are scope boundaries rather than validity flaws.

7.4 Metric Decoupling: Implications for Constrained Generation Tasks

The decoupling between structural validity (L2) and content fidelity (L0) has implications beyond legislative consolidation. In constrained generation tasks, formal validity can be achieved while semantic correctness fails. Similar decoupling likely appears in code generation (syntactically valid but semantically wrong), semantic parsing (valid structure, wrong meaning), and query generation (valid SQL, wrong logic). Therefore, evaluation of constrained generation should be multidimensional rather than single-metric.

7.5 Scope Boundaries and Future Work

While we considered 67 bills from the 17th Legislature (one jurisdiction) in the dataset used for our experiments, Italian legislative language and Akoma Ntoso implementations may differ from other parliaments. Future work should expand to additional jurisdictions, languages, and legislative traditions.

Furthermore, our results reflect January 2026 model versions. LLM capabilities evolve rapidly; our findings establish a methodology and baseline rather than permanent rankings. This allows for a framework that can be applied to repeat these tests in the future, assessing whether future models can improve upon the existing ones. Furthermore, as the outputs of LLMs are susceptible to changes in the prompt, a more thorough evaluation of different prompting strategies—including automatic optimization techniques—could be performed; smaller fine-tuned models may be more cost-effective than frontier LLMs for such ablations.

Our metrics assess similarity to ground truth, not independent legal validity. A model might produce output differing from ground truth while representing a legally valid alternative interpretation. This suggests that it would be useful to perform further analysis on the results by performing a manual annotation of the results for these legal criteria. A qualitative taxonomy of content errors—wrong target article, incorrect operation type, mishandled cascade amendments—would further guide prompt refinement.

In terms of context, several large bills exceed practical context or runtime limits, causing failures (especially for Kimi and DeepSeek). The underlying dataset is larger, but very large XML files are currently infeasible without chunking. Further work could try and tackle this task, perhaps using an agent-based approach to leverage multiple agents that could work on smaller fragments of the document.

In our experiments we report single runs per model-case pair with temperature=0.0. While this limits the stochastic nature of the LLM models by inhibiting probabilistic sampling, a better representation of the behaviour of the models could be achieved by repeating the test multiple times and measuring the mean and standard deviation of all scores. This approach, however, increases the time and cost of each experiment.

8 REPRODUCIBILITY AND TRANSPARENCY

We release the evaluation code (L0, BLEU-2, ROUGE-2, L1, L2), dataset manifest (67 bills, 23,416 amendments, official *DDL Messaggio*), per-case metrics for all complete outputs,

raw XML model outputs, and prompt templates (final version plus change history). These materials enable verification, extension to new models, analysis of error patterns, and adaptation to other legislatures. The repository link is provided in the Data Availability statement.

9 CONCLUSION

This paper presented a challenging task that is also hard for human beings with a high-level of legal drafting competence. In the past, this task was faced with NLP techniques and supported by semi-automatic tools. In this paper, we presented a systematic multi-metric evaluation of large language models for legislative consolidation of the bill, starting from the amendments, introducing a five-metric evaluation framework combining character fidelity (L0), bigram precision (BLEU-2), bigram overlap (ROUGE-2), semantic preservation (L1), and structural integrity (L2).

Our evaluation of six frontier LLMs on 67 Italian Senate bills comprising 23,416 amendments yields three principal findings:

(1) **Effective automation is achievable**: Claude Sonnet 4.5 achieves 76.9% character-level accuracy with 99.0% structural validity, demonstrating that LLMs can meaningfully automate legislative consolidation when the structure is well defined.
(2) **Structural validity alone is insufficient**: DeepSeek v3.2 produces structurally valid XML (L2=0.873) with fundamentally incorrect content (L0=0.210). Multi-metric evaluation is essential.
(3) **Error profiles are dominated by structural defects**: Missing `eId` attributes and malformed XML structure are common failure modes, motivating automated structural checks.

Beyond legislative drafting, the decoupling between structural validity and content fidelity generalises to other constrained generation tasks (e.g., code, semantic parsing, and query generation), where syntactic validity can coexist with semantic failure. Evaluations in these domains should similarly combine content, semantic, and structural metrics; structural validity alone is not a sufficient gate for legal AI deployment.

Data Availability: Dataset, evaluation code, and model outputs available at https://gitlab.com/CIRSFID/COGENTA.

ACKNOWLEDGMENTS

This project is conducted with the support of the European Commission funds within ERC HyperModeLex. Grant agreement ID: 101055185.

REFERENCES

[1] Monica Palmirani and Fabio Vitali. 2011. Akoma-Ntoso for Legal Documents. In Giovanni Sartor, Monica Palmirani, Enrico Francesconi, and Maria Angela Biasiotti (Eds.), *Legislative XML for the Semantic Web: Principles, Models, Standards for Document Management*. Springer, 75–100. DOI: https://doi.org/10.1007/978-94-007-1887-6_6. URL: https://link.springer.com/book/10.1007/978-94-007-1887-6.

[2] Adam Wyner and Wim Peters. 2011. On Rule Extraction from Regulations. In *Proceedings of the 24th International Conference on Legal Knowledge and Information Systems (JURIX 2011)*. IOS Press, 113–122. DOI: https://doi.org/10.3233/978-1-60750-981-3-113. URL: https://ebooks.iospress.nl/publication/6692.

[3] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. In *NeurIPS 2021 Datasets and Benchmarks Track*. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/6ea9ab1baa0efb9e19094440c317e21b-Abstract-round1.html.

[4] Michael A. Bommarito II and Daniel M. Katz. 2022. GPT Takes the Bar Exam. arXiv:2212.14402. DOI: https://doi.org/10.48550/arXiv.2212.14402. URL: https://michaelbommarito.com/publications/2022-gpt-takes-the-bar-exam/.

[5] Jonathan H. Choi, Kristin E. Hickman, Amy B. Monahan, and Daniel B. Schwarcz. 2022. ChatGPT Goes to Law School. *Journal of Legal Education* 71:387. URL: https://scholarship.law.umn.edu/faculty_articles/1044/.

[6] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of EMNLP 2019*. 3982–3992. DOI: https://doi.org/10.18653/v1/D19-1410. URL: https://aclanthology.org/D19-1410/.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*. 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*. 311–318. DOI: https://doi.org/10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040/.

[9] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. 74–81. URL: https://aclanthology.org/W04-1013/.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of NeurIPS 2017*. 5998–6008. DOI: https://doi.org/10.48550/arXiv.1706.03762. URL: https://arxiv.org/abs/1706.03762.

[11] Alex L. Zhang, Tim Kraska, and Omar Khattab. 2025. Recursive Language Models. arXiv:2512.24601. DOI: https://doi.org/10.48550/arXiv.2512.24601. URL: https://arxiv.org/abs/2512.24601.

[12] Matias Etcheverry, Thibaud Real, and Pauline Chavallard. 2024. Algorithm for Automatic Legislative Text Consolidation. In *Proceedings of the Natural Legal Language Processing Workshop 2024*. 166–175. DOI: https://doi.org/10.18653/v1/2024.nllp-1.13. URL: https://aclanthology.org/2024.nllp-1.13/.

[13] Inter-Parliamentary Union (IPU). *World e-Parliament Report 2024*. Inter-Parliamentary Union, 2024. Available at: https://www.ipu.org/resources/publications/reports/2024-10/world-e-parliament-report-2024.

[14] Senato della Repubblica. *The challenges of artificial intelligence: approaches to their regulation and parliamentary use*. Focus dossier,

Senate of the Republic, Impact Assessment Office, 2025. Available at: https://www.senato.it/application/xmanager/projects/leg19/attachments/documento/files/000/112/797/DA31_Focus_Artificial_intelligence.pdf.

[15] Bar-Siman-Tov, I. *Legislatures and legislation in the age of artificial intelligence*. The Theory and Practice of Legislation, vol. 13, no. 3, 2025, pp. 267–291. https://doi.org/10.1080/20508840.2025.2590971.

[16] Bonomi, F., Lupo, N., and Piccirilli, G. *Logical and Procedural Rules for Parliamentary Amendments, in Light of the Italian Experience, from Bentham to AI*. The Theory and Practice of Legislation (TPLeg), 2025, p. 1.

[17] Y. Li and B. Liu, "A Normalized Levenshtein Distance Metric," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1091-1095, June 2007.